Uniwersytet
Ekonomiczny
w Katowicach

Analiza danych jakościowych

Bogna Zacny

zima 2019/2020

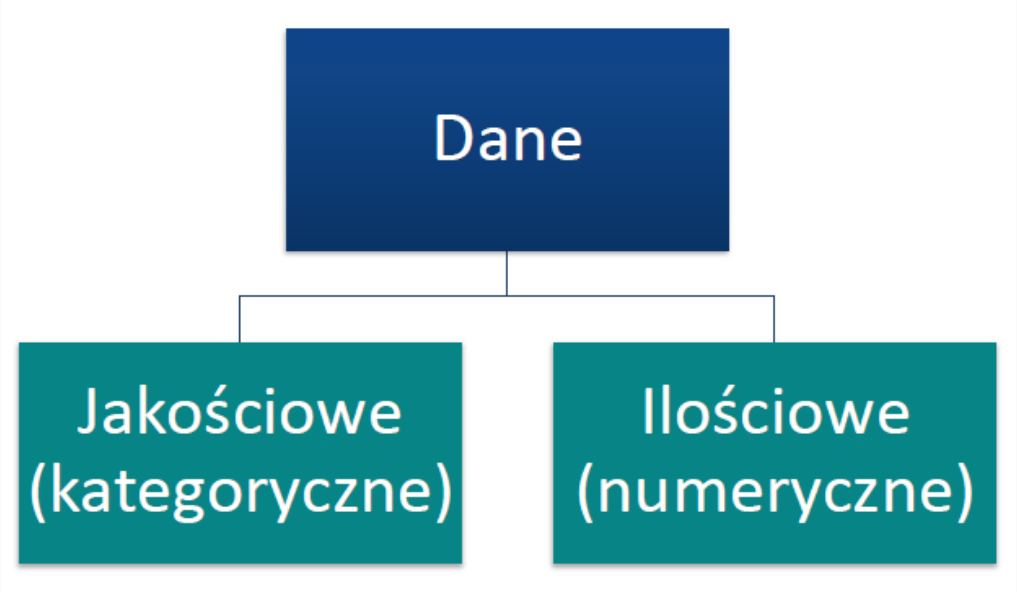
Wydział Informatyki i Komunikacji
Katedra Inżynierii Wiedzy

Agenda

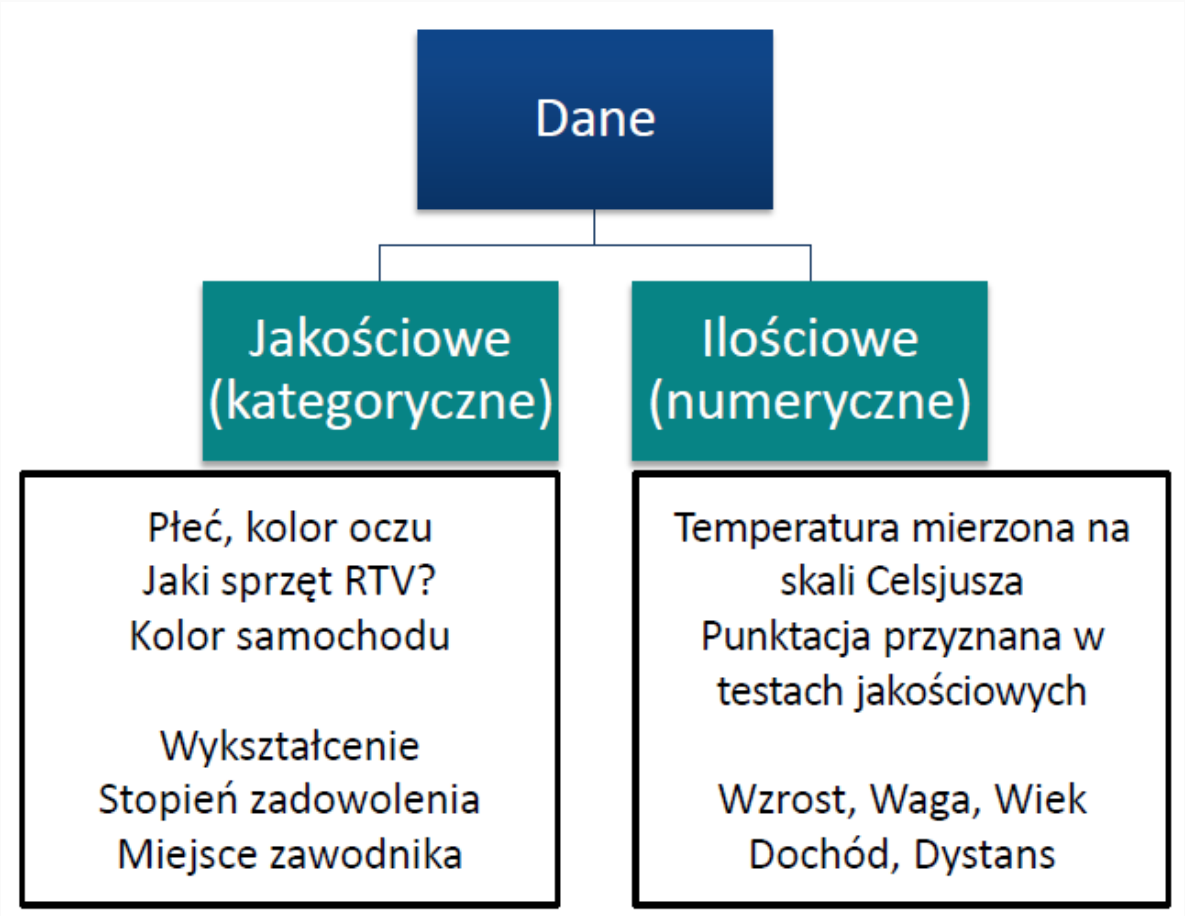
1. Dane
2. Tablice kontyngencji
3. Wykres mozaikowy
4. Analiza korespondencji

Dane

Dane - skale



Dane - skale


Dane
Jakościowe (kategoryczne)
Ilościowe (numeryczne)
Płeć, kolor oczu
Jaki sprzęt RTV?
Kolor samochodu
Wykształcenie
Stopień zadowolenia
Miejsce zawodnika
Temperatura mierzona na skali Celsjusza
Punktacja przyznana w testach jakościowych
Wzrost, Waga, Wiek
Dochód, Dystans

Dane - Titanic

Sex	Pclass	Survived
male	III	N
female	I	Y
female	III	Y
female	I	Y
male	III	N
male	III	N
male	I	N
male	III	N
female	III	Y
female	II	Y

Dane - Titanic

Sex	Pclass	Survived
103	3	0
105	1	1
105	3	1
105	1	1
103	3	0
103	3	0
103	1	0
103	3	0
105	3	1
105	2	1

Dane - Titanic

Sex	Pclass	Survived
0	3	0
1	1	1
1	3	1
1	1	1
0	3	0
0	3	0
0	1	0
0	3	0
1	3	1
1	2	1

Dane - Titanic

Sex_female	Sex_male	Pclass_1	Pclass_2	Pclass_3	Survived_0	Survived_1
0	1	0	0	1	1	0
1	0	1	0	0	0	1
1	0	0	0	1	0	1
1	0	1	0	0	0	1
0	1	0	0	1	1	0
0	1	0	0	1	1	0
0	1	1	0	0	1	0
0	1	0	0	1	1	0
1	0	0	0	1	0	1
1	0	0	1	0	0	1

Dane - Kto słucha muzyki klasycznej?

Wyksztalcenie	Klasyka	Wiek
Niskie	Tak	Młody
Wysokie	Nie	Młody
Wysokie	Nie	Stary
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary
Niskie	Tak	Młody
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary

Dane - Kto słucha muzyki klasycznej?

Wyksztalcenie	Klasyka	Wiek
Niskie	Tak	Młody
Wysokie	Nie	Młody
Wysokie	Nie	Stary
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary
Niskie	Tak	Młody
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary

Dane - Kto słucha muzyki klasycznej?

Wyksztalcenie	Klasyka	Wiek
Niskie	Tak	Młody
Wysokie	Nie	Młody
Wysokie	Nie	Stary
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary
Niskie	Tak	Młody
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary

	Nie	Tak
Młody	1	2
Stary	5	2

Dane - Kto słucha muzyki klasycznej?

Wyksztalcenie	Klasyka	Wiek
Niskie	Tak	Młody
Wysokie	Nie	Młody
Wysokie	Nie	Stary
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary
Niskie	Tak	Młody
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary

	Nie	Tak
Młody	1	2
Stary	5	2

	Nie	Tak
Niskie	4	2
Wysokie	2	2

Dane - Kto słucha muzyki klasycznej?

Wyksztalcenie	Klasyka	Wiek
Niskie	Tak	Młody
Wysokie	Nie	Młody
Wysokie	Nie	Stary
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary
Niskie	Tak	Młody
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary

Dane - Kto słucha muzyki klasycznej?

Wyksztalcenie	Klasyka	Wiek
Niskie	Tak	Młody
Wysokie	Nie	Młody
Wysokie	Nie	Stary
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary
Niskie	Tak	Młody
Wysokie	Tak	Stary
Niskie	Nie	Stary
Niskie	Nie	Stary

Wyksztalcenie	Klasyka	Wiek	Freq
Niskie	Nie	Młody	0
Wysokie	Nie	Młody	1
Niskie	Tak	Młody	2
Wysokie	Tak	Młody	0
Niskie	Nie	Stary	4
Wysokie	Nie	Stary	1
Niskie	Tak	Stary	0
Wysokie	Tak	Stary	2

Dane - Kto słucha muzyki klasycznej?

Wyksztalcenie	Klasyka	Wiek	Freq
Niskie	Nie	Młody	290
Wysokie	Nie	Młody	406
Niskie	Tak	Młody	110
Wysokie	Tak	Młody	194
Niskie	Nie	Stary	730
Wysokie	Nie	Stary	190
Niskie	Tak	Stary	170
Wysokie	Tak	Stary	210

Tablice kontyngencji

Dwuwymiarowe tablice kontyngencji 2x2

X	Y		
	y_1	y_2	$n_{i\bullet}$
x_1	n_{11}	n_{12}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$n_{2\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Dwuwymiarowe tablice kontyngencji *IxJ*

X	Y			
	y_1	...	y_J	$n_{i\bullet}$
x_1	n_{11}	...	n_{1J}	$n_{1\bullet}$
⋮	⋮	⋮	⋮	⋮
x_I	n_{I1}	...	n_{IJ}	$n_{I\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	...	$n_{\bullet J}$	n

Wielowymiarowe tablice kontyngencji

Z	z_1				...	z_k				
	X	Y				X	Y			
		y_{11}	...	y_{J1}	$n_{i\bullet 1}$		y_{1k}	...	y_{Jk}	$n_{i\bullet k}$
	x_{11}	n_{111}	...	n_{1J1}	$n_{1\bullet 1}$	x_{1k}	n_{11k}	...	n_{1Jk}	$n_{1\bullet k}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	x_{I1}	n_{I11}	...	n_{IJ1}	$n_{I\bullet 1}$	x_{Ik}	n_{I1k}	...	n_{IJk}	$n_{I\bullet k}$
	$n_{\bullet j1}$	$n_{\bullet 11}$	...	$n_{\bullet J1}$	n_1	$n_{\bullet jk}$	$n_{\bullet 1k}$	...	$n_{\bullet Jk}$	n_k
										n

Mierniki siły zależności

- Współczynnik chi-kwadrat: $\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$

gdzie: $m_{ij} = \frac{n_{\cdot i} * n_{j\cdot}}{n}$

- H_0: zmienne kolumnowa i wierszowa tablicy kontyngencji sa niezależne
- H_1: zmienne kolumnowa i wierszowa tablicy kontyngencji sa zależne

Mierniki siły zależności

- Współczynnik fi (2×2): $\phi = \sqrt{\frac{\chi^2}{n}}$
- Współczynnik kontyngencji: $C = \sqrt{\frac{\chi^2}{\chi^2+n}}$
- Współczynnik V Cramera: $V = \sqrt{\frac{\chi^2}{n*min(I-1,J-1)}}$

Czy zależność istnieje?

	tak	nie
m	50	0
s	0	50

```
##
##  Pearson's Chi-squared test
##
## data:  kontyng
## X-squared = 100, df = 1, p-value < 2.2e-16
```

Czy zależność istnieje?

	tak	nie
m	50	0
s	0	50

```
##
##  Pearson's Chi-squared test
##
## data:  kontyng
## X-squared = 100, df = 1, p-value < 2.2e-16
```

	tak	nie
m	25	25
s	25	25

```
##
##  Pearson's Chi-squared test
##
## data:  kontyng
## X-squared = 0, df = 1, p-value = 1
```

Czy zależność istnieje?

	tak	nie
m	45	5
s	5	45

```
##
##  Pearson's Chi-squared test
##
## data:  kontyng
## X-squared = 64, df = 1, p-value = 1.244e-15
```

Czy zależność istnieje?

	tak	nie
m	45	5
s	5	45

```
##
##  Pearson's Chi-squared test
##
## data:  kontyng
## X-squared = 64, df = 1, p-value = 1.244e-15
```

	tak	nie
m	35	15
s	15	35

```
##
##  Pearson's Chi-squared test
##
## data:  kontyng
## X-squared = 16, df = 1, p-value = 6.334e-05
```

Czy zależność istnieje?

	tak	nie
m	45	5
s	0	0

```
##
##  Pearson's Chi-squared test
##
## data:  kontyng
## X-squared = NaN, df = 1, p-value = NA
```

Czy zależność istnieje?

	tak	nie
m	45	25
s	5	5

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  kontyng
## X-squared = 0.27429, df = 1, p-value = 0.6005
```

Jak silna jest zależność?

	tak	nie
m	50	0
s	0	50

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 138.63  1        0
## Pearson          100.00  1        0
##
## Phi-Coefficient   : 1
## Contingency Coeff.: 0.707
## Cramer's V        : 1
```

Jak silna jest zależność?

	tak	nie
m	25	25
s	25	25

```
##                   X^2 df P(> X^2)
## Likelihood Ratio    0  1        1
## Pearson             0  1        1
##
## Phi-Coefficient   : 0
## Contingency Coeff.: 0
## Cramer's V        : 0
```

Jak silna jest zależność?

	tak	nie
m	35	15
s	15	35

```
##                      X^2 df   P(> X^2)
## Likelihood Ratio 16.457  1 4.9777e-05
## Pearson          16.000  1 6.3342e-05
##
## Phi-Coefficient   : 0.4
## Contingency Coeff.: 0.371
## Cramer's V        : 0.4
```

Jak silna jest zależność?

	tak	trochę	nie
podst	35	10	9
śred	15	38	7
wyz	2	4	44

```
##                      X^2 df P(> X^2)
## Likelihood Ratio 112.53  4        0
## Pearson          114.54  4        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.641
## Cramer's V        : 0.591
```

Kto słucha klasyki? (Czy zależność istnieje?)

	Nie	Tak
Niskie	1020	280
Wysokie	596	404

```
chisq.test(table(mozaikowy$Wyksztalcenie, mozaikowy$Klasyka))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(mozaikowy$Wyksztalcenie, mozaikowy$Klasyka)
## X-squared = 95.333, df = 1, p-value < 2.2e-16
```

Kto słucha klasyki? (Czy zależność istnieje?)

	Nie	Tak
Młody	696	304
Stary	920	380

```
chisq.test(table(mozaikowy$Wiek, mozaikowy$Klasyka))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(mozaikowy$Wiek, mozaikowy$Klasyka)
## X-squared = 0.31597, df = 1, p-value = 0.574
```

Kto słucha klasyki? (Jak silna jest zależność)

```
assocstats(table(mozaikowy$Wyksztalcenie,
                 mozaikowy$Klasyka))
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 95.923  1        0
## Pearson          96.234  1        0
##
## Phi-Coefficient   : 0.205
## Contingency Coeff.: 0.2
## Cramer's V        : 0.205
```

Siła związku - Kto słucha klasyki?

```
assocstats(table(mozaikowy$Wiek,
                 mozaikowy$Klasyka))
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 0.36946  1  0.54330
## Pearson          0.36981  1  0.54311
##
## Phi-Coefficient   : 0.013
## Contingency Coeff.: 0.013
## Cramer's V        : 0.013
```

Wykres mozaikowy

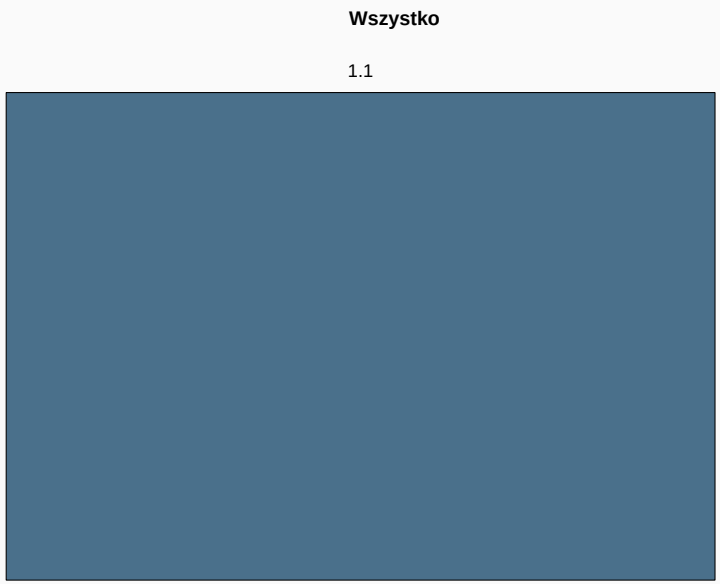
Wykres mozaikowy - opis

Wykres mozaikowy stosowany aby prezentować kilkupoziomowe hierarchie, które na kolejnych poziomach dzielone są tymi samymi czynnikami.

	Wykształczenie			
	Niskie		Wysokie	
	Klasyka		Klasyka	
Wiek	Nie	Tak	Nie	Tak
Młody	290	110	406	194
Stary	730	170	190	210

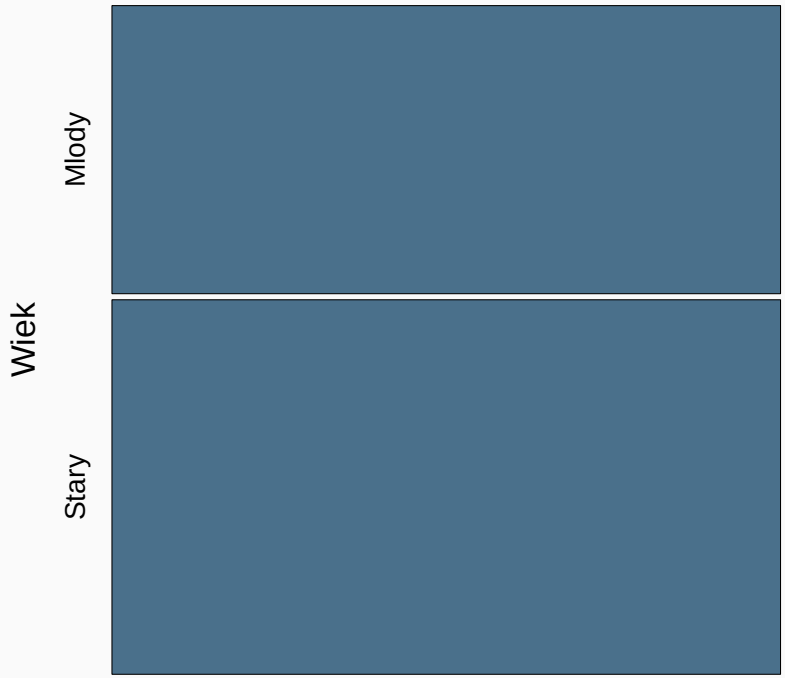
Wykres mozaikowy - 0 wymiarów

```
## [1] 2300
```



Wykres mozaikowy - 1 wymiar

Wiek	Freq
Młody	43.47826
Stary	56.52174



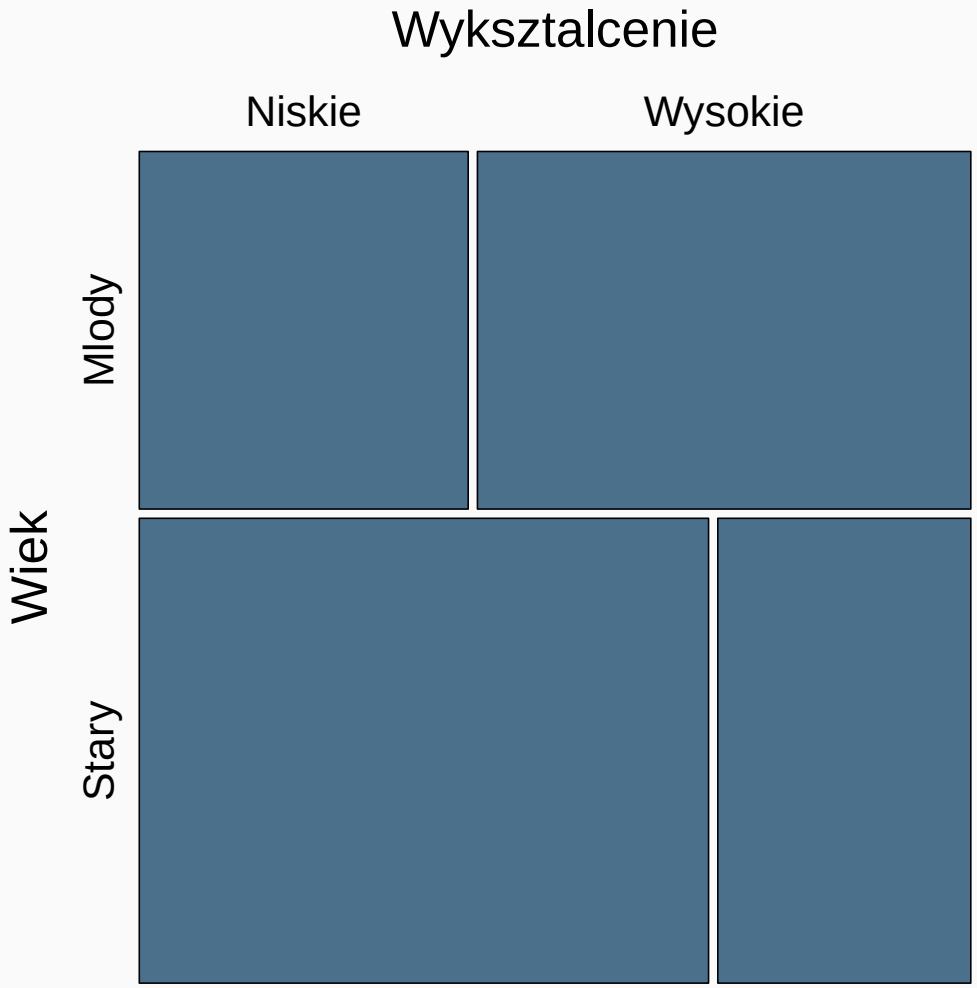
Wykres mozaikowy - 1 wymiar

```
mosaic(~ Wiek, mozaikowy,
       gp = gpar(fill = c("skyblue4")),
       labeling = labeling_cboxed(tl_labels = TRUE,
                                  boxes = F,
                                  clip = TRUE,
                                  pos_labels = "center",
                                  offset_varnames = c(left = 0.2),
                                  offset_labels = c(left = 1.5),
                                  gp_labels=gpar(fontsize=20),
                                  gp_varnames=gpar(fontsize=24)),
       margins=unit(1, "lines"))
```

Wykres mozaikowy - 2 wymiary

	Niskie	Wysokie
Młody	17.39130	26.08696
Stary	39.13043	17.39130

Wykres mozaikowy - 2 wymiary



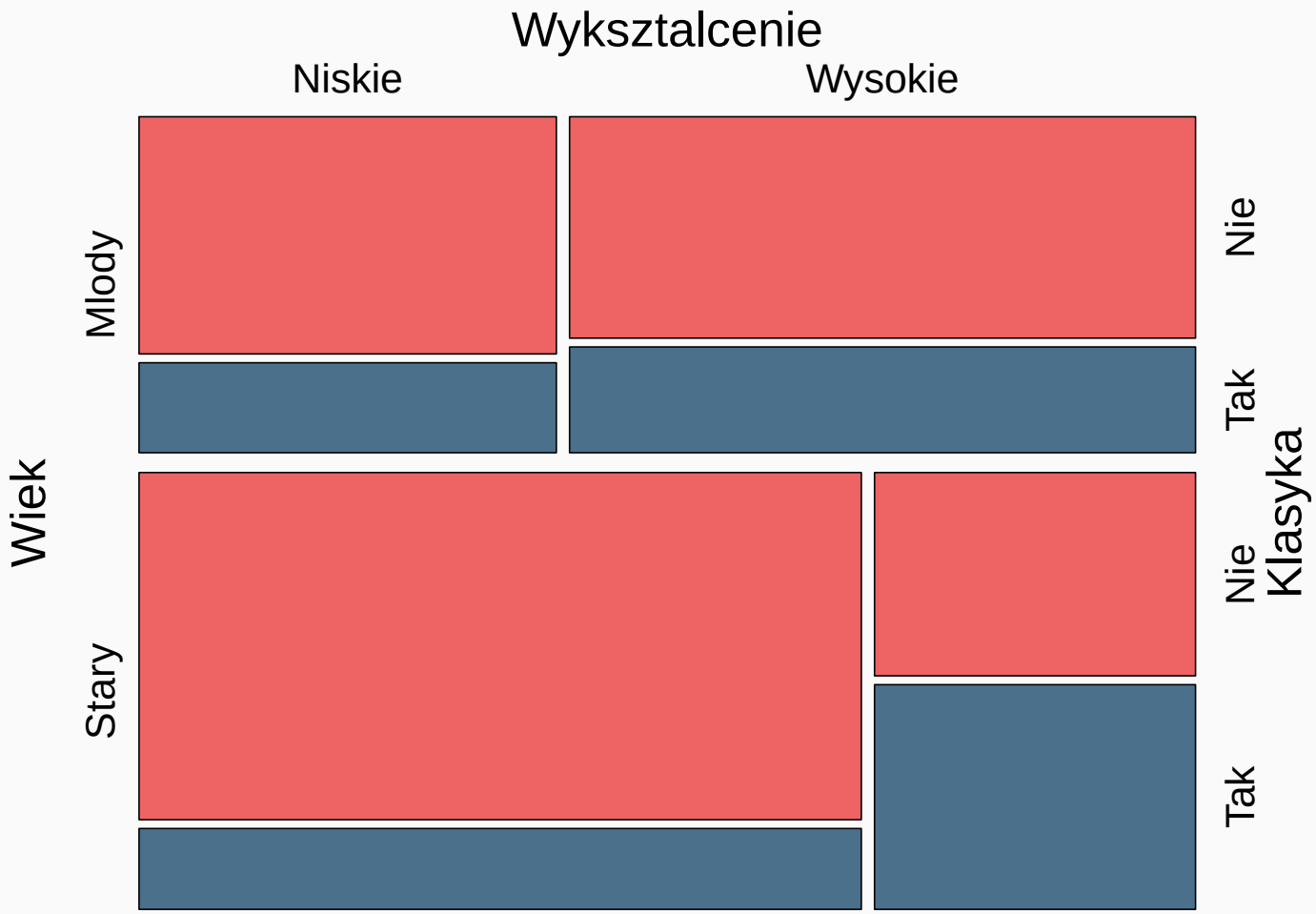
Wykres mozaikowy - 2 wymiary

```
mosaic(~ Wiek + Wyksztalcenie, mozaikowy,
       gp = gpar(fill = c("skyblue4")),
       labeling = labeling_cboxed(tl_labels = TRUE,
          boxes = F,
          clip = TRUE,
          pos_labels = "center",
          offset_varnames = c(left = -0.2),
          offset_labels = c(left = 1, 1),
          gp_labels=gpar(fontsize=20),
          gp_varnames=gpar(fontsize=24)),
       margins=unit(c(7,1,1,3), "lines"))
```

Wykres mozaikowy - 3 wymiary

Wiek	Wyksztalcenie	Klasyka	Freq
Młody	Niskie	Nie	12.608696
Stary	Niskie	Nie	31.739130
Młody	Wysokie	Nie	17.652174
Stary	Wysokie	Nie	8.260870
Młody	Niskie	Tak	4.782609
Stary	Niskie	Tak	7.391304
Młody	Wysokie	Tak	8.434783
Stary	Wysokie	Tak	9.130435

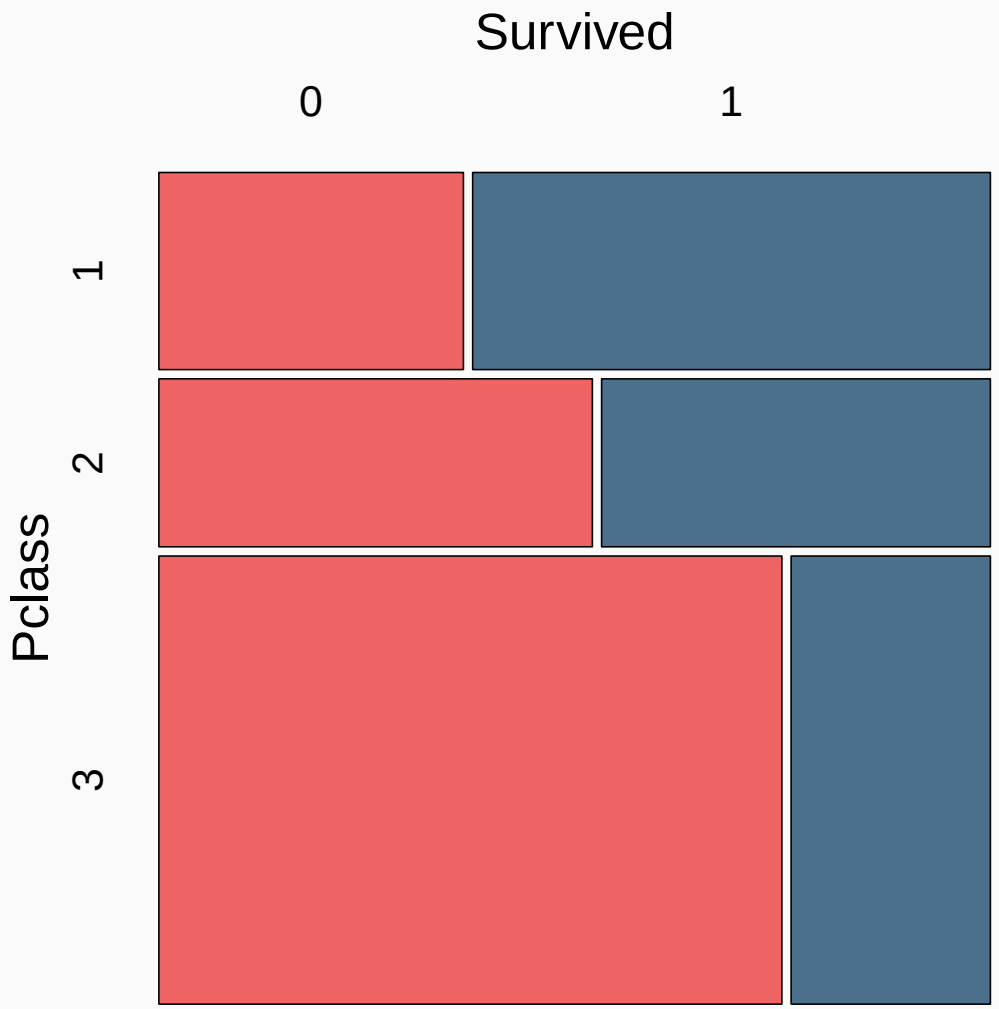
Wykres mozaikowy - 3 wymiary



Wykres mozaikowy - 3 wymiary

```
mosaic(~ Wiek + Wyksztalcenie + Klasyka, mozaikowy,
       gp = gpar(fill = c("indianred2", "skyblue4")),
       labeling = labeling_border(
                  boxes = F,
                  clip = TRUE,
                  pos_labels = "center",
                  offset_varnames = c(left = -0.1,
                                      top = -0.5,
                                      right = -0.5),
                  offset_labels = c(left = 1, 1),
                  gp_labels=gpar(fontsize=20),
                  gp_varnames=gpar(fontsize=24)),
       margins=unit(c(7,7,1,7), "lines"))
```

Wykres mozaikowy - przykład



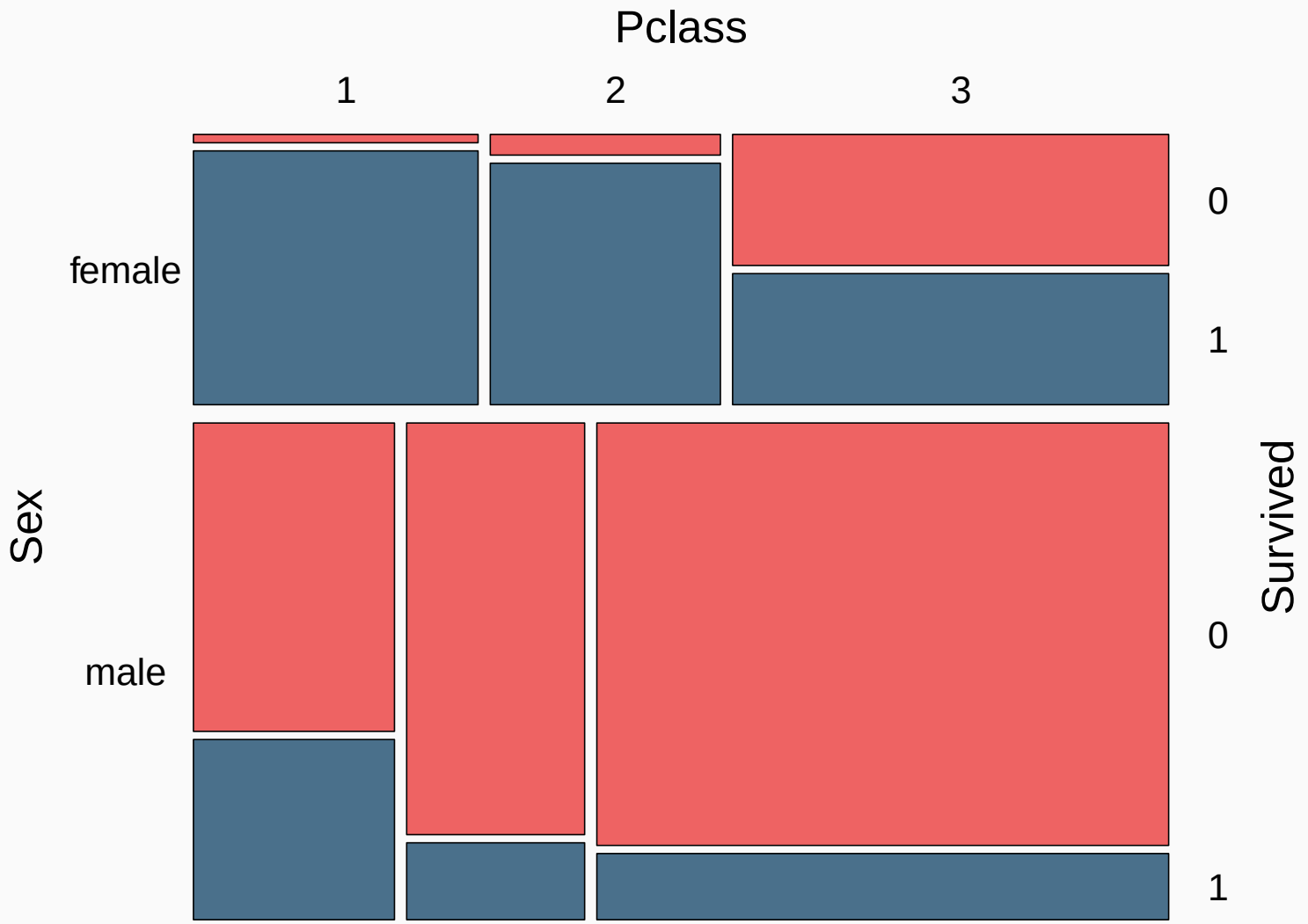
Wykres mozaikowy - przykład

```
mosaic(~ Pclass + Survived, dane_Titanic,
       gp = gpar(fill = c("indianred2", "skyblue4")),
       labeling = labeling_border(
                  boxes = F,
                  clip = TRUE,
                  pos_labels = "center",
                  offset_varnames = c(left = 0.1,
                                      top = 0.3),
                  offset_labels = c(left = 2, 2),
                  gp_labels=gpar(fontsize=20),
                  gp_varnames=gpar(fontsize=24)),
       margins=unit(c(7,1,1,7), "lines"))
```

Wykres mozaikowy - przykład

Wykres mozaikowy, przedstawia udział osób z różnych klas w rozbiciu na przeżycie. Pole każdego prostokąta odpowiada udziałowi określonej grupy klasa/przeżycie w populacji, wysokość prostokąta odpowiada udziałowi osób z danej klasy a szerokość udziałowi osób, które przeżyły lub nie.

Wykres mozaikowy - przykład



Wykres mozaikowy - przykład

```
mosaic(~ Sex + Pclass + Survived , dane_Titanic,
       gp = gpar(fill = c("indianred2", "skyblue4")),
       labeling= labeling_border(
                    rot_labels = c(0,0,90,0),
                    just_labels = c("left", "center",
                                    "center", "center"),
                    pos_labels = "center",
                    offset_varnames = c(left = 1.1),
                    offset_labels = c(left = 0.5),
                    gp_labels=gpar(fontsize=20),
                    gp_varnames=gpar(fontsize=24)),
       margins=unit(c(5,7,1,7), "lines"))
```

Analiza korespondencji

Analiza korespondencji - opis

Celem analizy korespondencji jest graficzne przedstawienie wyników analizy zależności zmiennych wielokategorialnych (liczba kategorii większa od 2) na mapie percepcji w niskowymiarowej przestrzeni

Analiza korespondencji - obowiązki domowe

```
## [1] "Wife"        "Alternating" "Husband"     "Jointly"
```

```
##  [1] "Laundry"    "Main_meal"  "Dinner"     "Breakfeast" "Tidying"
##  [6] "Dishes"     "Shopping"   "Official"   "Driving"    "Finances"
## [11] "Insurance"  "Repairs"    "Holidays"
```

```
## # A tibble: 6 x 4
##    Wife Alternating Husband Jointly
##   <int>       <int>   <int>   <int>
## 1   156          14       2       4
## 2   124          20       5       4
## 3    77          11       7      13
## 4    82          36      15       7
## 5    53          11       1      57
## 6    32          24       4      53
```

Analiza korespondencji - obowiązki domowe (m1)

	Wife	Husband	Jointly
Laundry	156	2	4
Repairs	0	160	2
Holidays	0	6	153

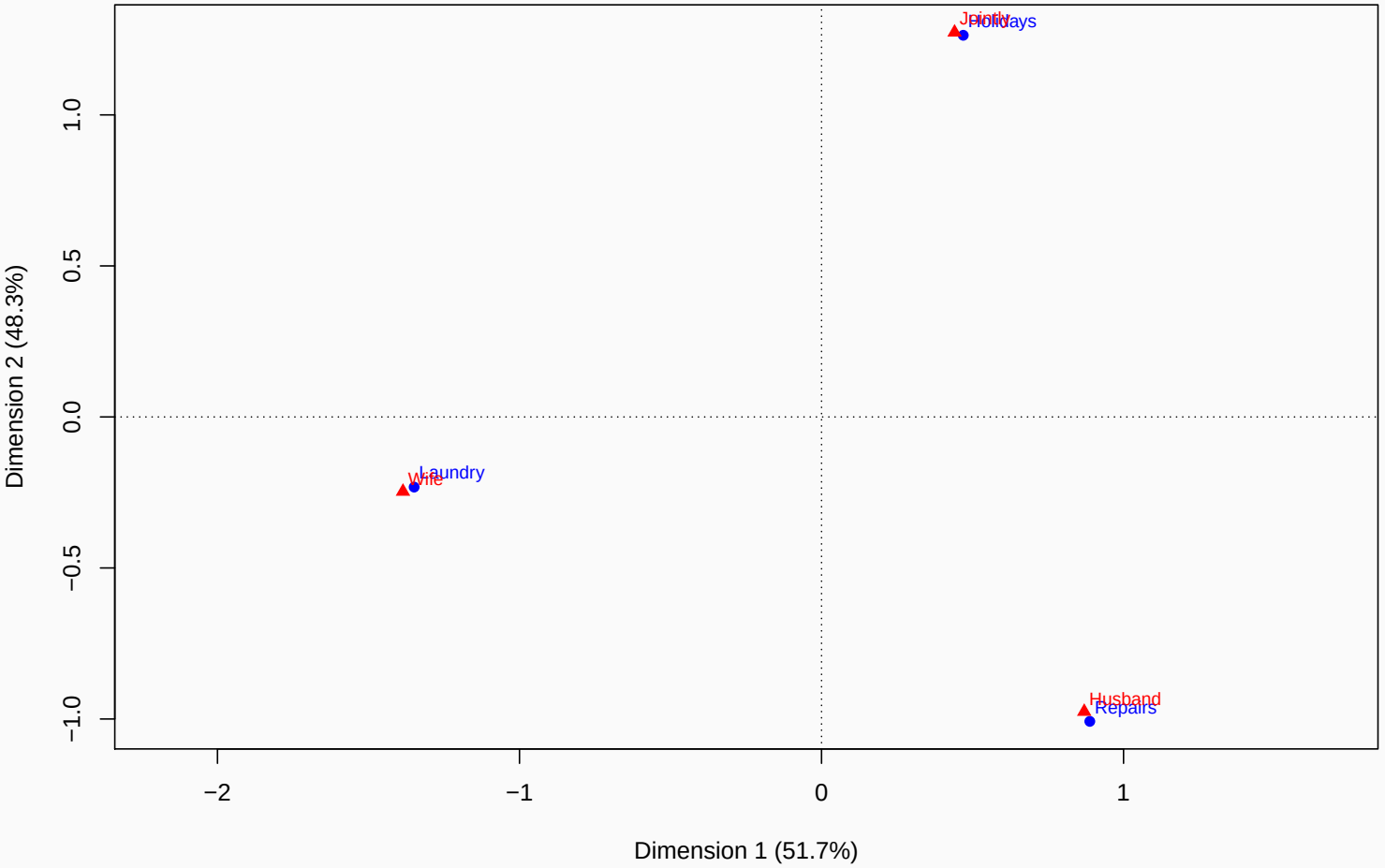
Analiza korespondencji - obowiązki domowe (m1)

```
library("ca")
wynik <- ca(domowe_maly)
```

Analiza korespondencji - obowiązki domowe (m1)

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.947364  51.7  51.7  ************
##  2      0.884446  48.3 100.0  ************
##         -------- -----
##  Total: 1.831810 100.0
##
##
## Rows:
##      name   mass  qlt  inr     k=1 cor ctr     k=2 cor ct
## 1 | Lndr |  335 1000  343 | -1349 971 644 |  -232  29  2
## 2 | Rprs |  335 1000  330 |   888 437 279 | -1008 563 38
## 3 | Hldy |  329 1000  327 |   469 121  77 |  1264 879 59
```

Analiza korespondencji - obowiązki domowe (m1)



Analiza korespondencji - obowiązki domowe (m2)

	Wife	Husband	Jointly
Laundry	156	2	4
Repairs	0	160	2
Dishes	32	4	53

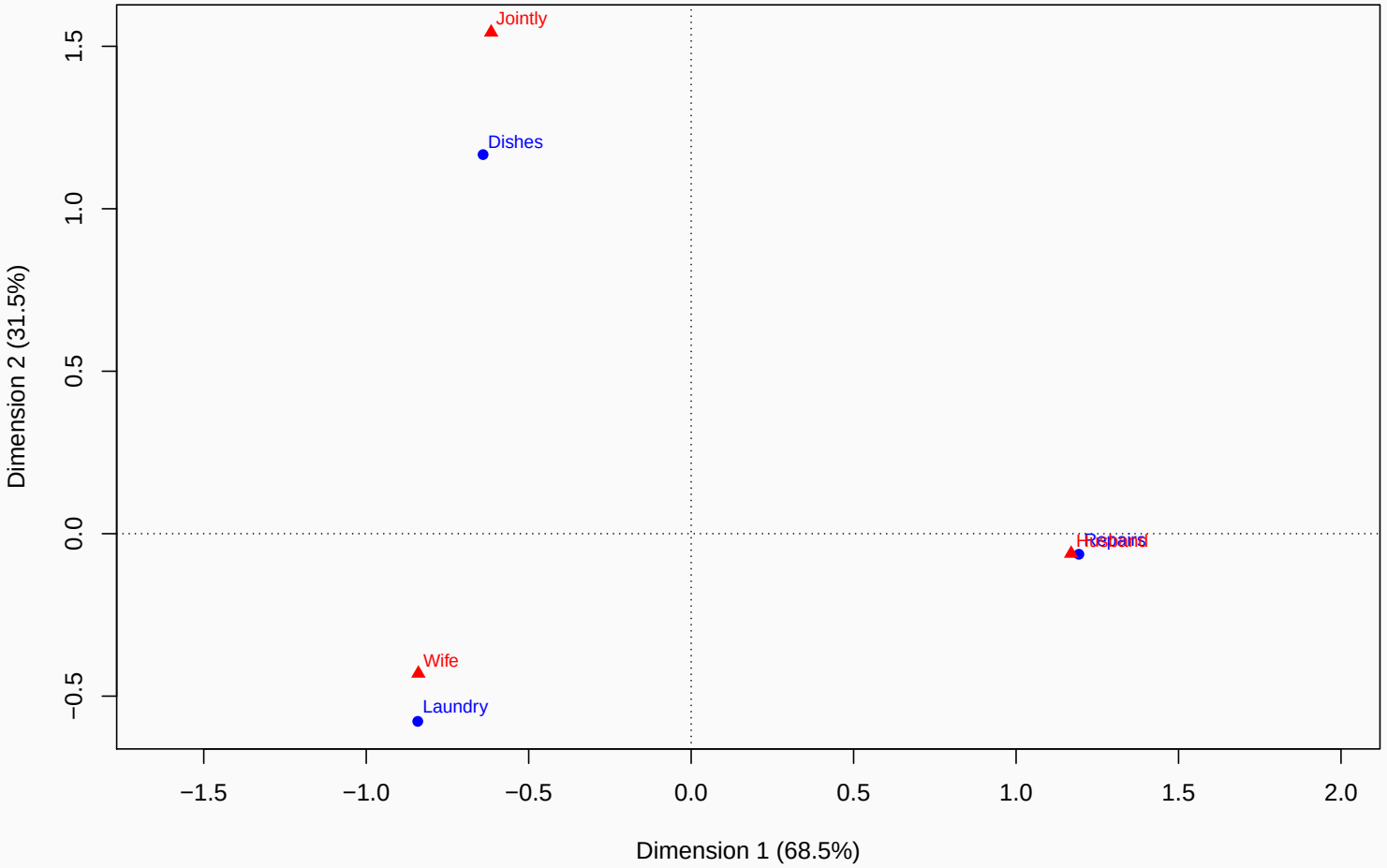
Analiza korespondencji - obowiązki domowe (m2)

```
library("ca")
wynik <- ca(domowe_maly)
```

Analiza korespondencji - obowiązki domowe (m2)

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.924669  68.5  68.5  ******************
##  2      0.425811  31.5 100.0  ********
##         -------- -----
##  Total: 1.350480 100.0
##
##
## Rows:
##      name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | Lndr |  392 1000  303 | -841 680 300 | -578 320 307
## 2 | Rprs |  392 1000  415 | 1193 997 604 |  -63   3   4
## 3 | Dshs |  215 1000  283 | -640 232  96 | 1167 768 689
```

Analiza korespondencji - obowiązki domowe (m2)



Analiza korespondencji - obowiązki domowe (m3)

	Wife	Husband	Jointly
Laundry	156	2	4
Repairs	0	160	2
Finances	13	21	66

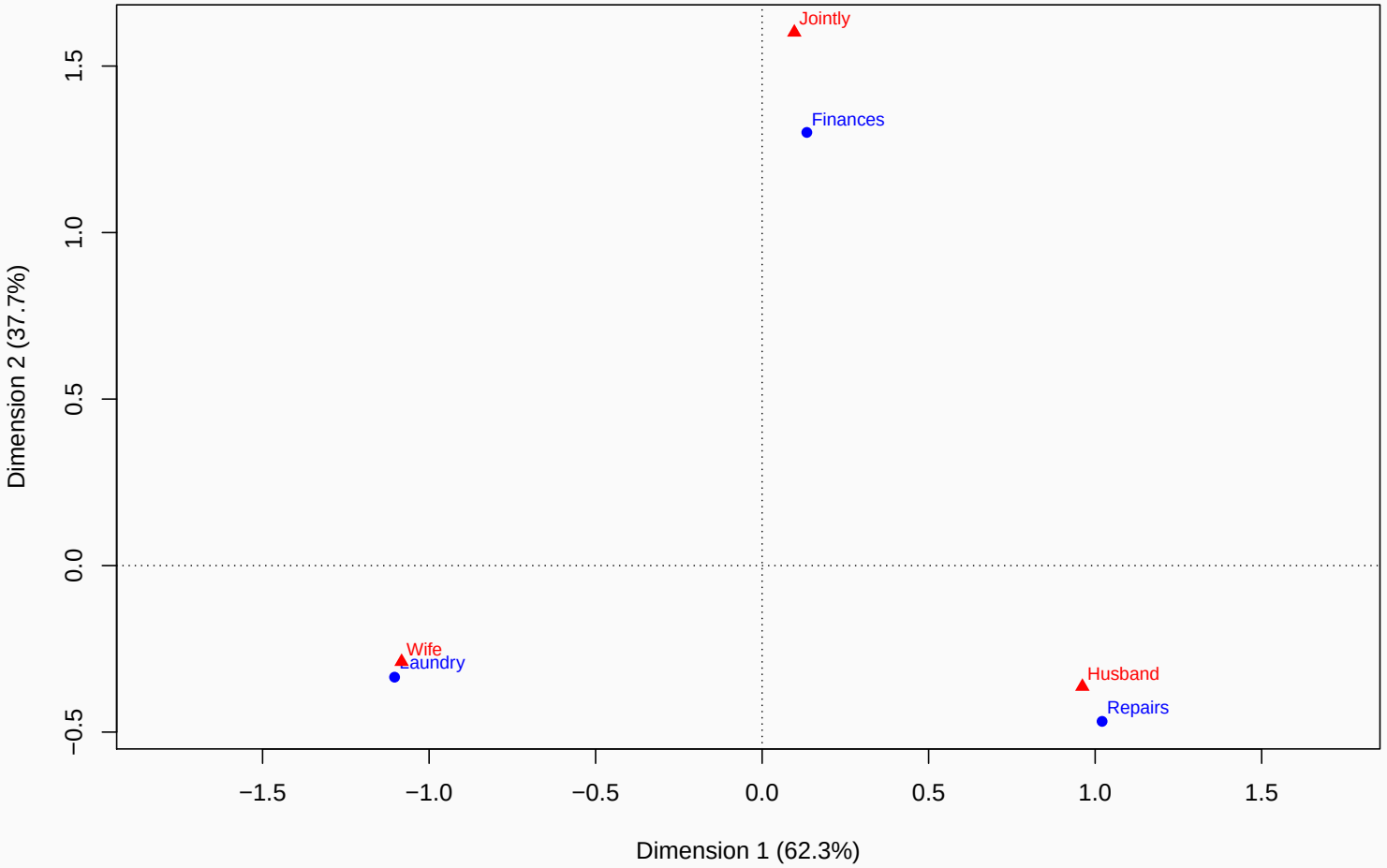
Analiza korespondencji - obowiązki domowe (m3)

```
library("ca")
wynik <- ca(domowe_maly)
```

Analiza korespondencji - obowiązki domowe (m3)

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.867844  62.3  62.3  ****************
##  2      0.525516  37.7 100.0  *********
##         -------- -----
##  Total: 1.393360 100.0
##
##
## Rows:
##      name   mass  qlt  inr     k=1 cor ctr    k=2 cor ctr
## 1 | Lndr |  382 1000  365 | -1104 916 536 | -335  84  82
## 2 | Rprs |  382 1000  346 |  1021 826 459 | -468 174 159
## 3 | Fnnc |  236 1000  289 |   134  11   5 | 1301 989 759
```

Analiza korespondencji - obowiązki domowe (m3)



Analiza korespondencji - obowiązki domowe (m4)

	Wife	Husband	Jointly
Laundry	156	2	4
Repairs	0	160	2
Official	12	23	15

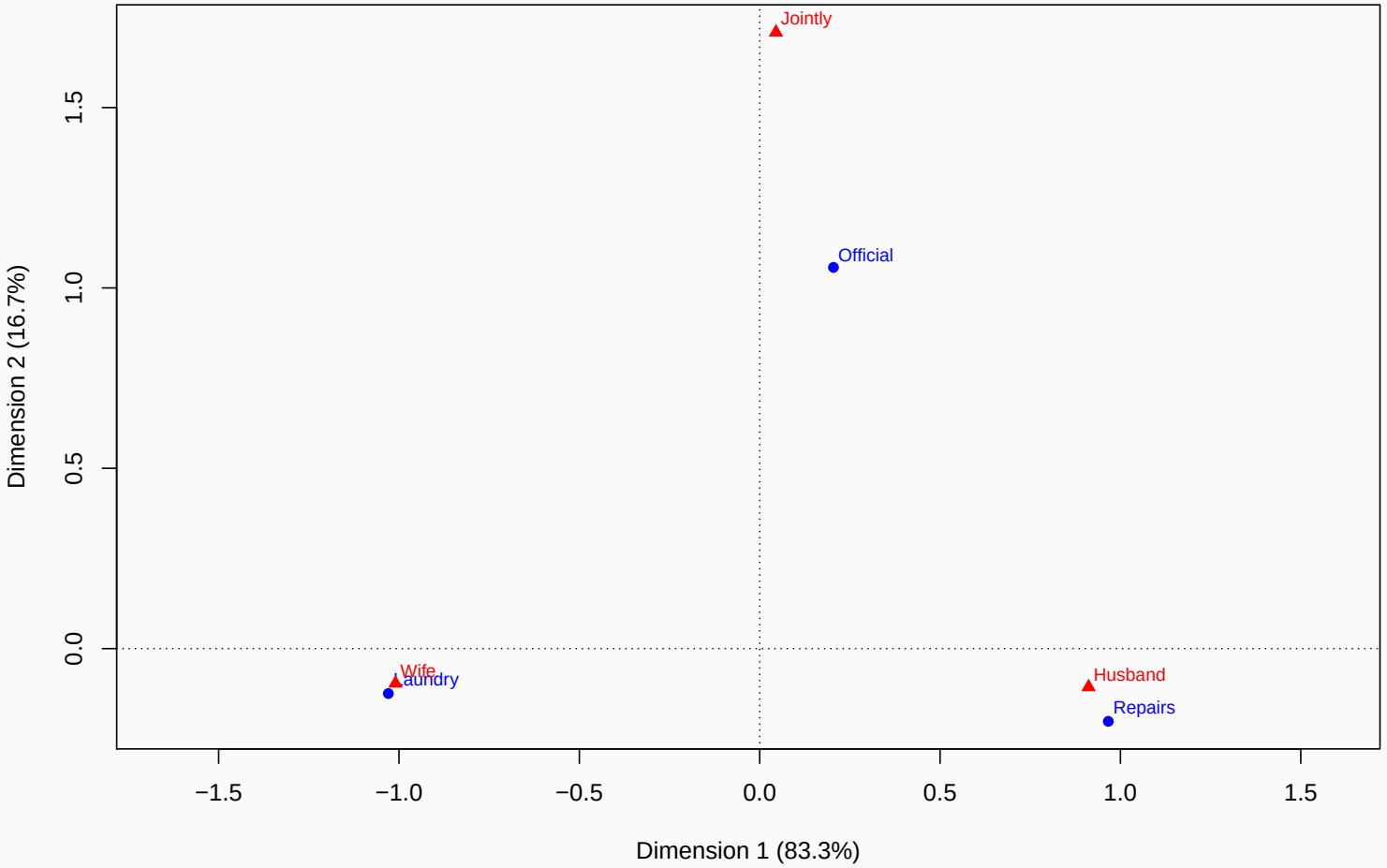
Analiza korespondencji - obowiązki domowe (m4)

```
library("ca")
wynik <- ca(domowe_maly)
```

Analiza korespondencji - obowiązki domowe (m4)

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %    cum%   scree plot
##  1      0.869322  83.3  83.3   **********************
##  2      0.173676  16.7 100.0   ****
##         -------- -----
##  Total: 1.042998 100.0
##
##
## Rows:
##      name   mass  qlt  inr     k=1 cor ctr    k=2 cor ctr
## 1 | Lndr |  433 1000  447 | -1030 986 528 | -125  14  39
## 2 | Rprs |  433 1000  405 |   966 958 465 | -202  42 101
## 3 | Offc |  134 1000  149 |   204  36   6 | 1057 964 860
```

Analiza korespondencji - obowiązki domowe (m4)



Analiza korespondencji - obowiązki domowe

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.542889  48.7  48.7  ************
##  2      0.445003  39.9  88.6  **********
##  3      0.127048  11.4 100.0  ***
##         -------- -----
##  Total: 1.114940 100.0
##
##
## Rows:
##      name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1  | Lndr |  101  925  120 | -992 740 183 | -495 185  56
## 2  | Mn_m |   88  974   81 | -876 742 124 | -490 232  47
```

Analiza korespondencji - obowiązki domowe